AFRL-OSR-VA-TR-2013-0560

SOCIAL DATA ANALYSIS BY NON LINEAR IMBEDDING

Zucker, Steven, W.
Yale University
51 Prospect St.
New Haven, CT

20-09-2013
Final Report

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 20-09-2013	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 15-12-2008 - 18--09--2013.

4. TITLE AND SUBTITLE
SOCIAL DATA ANALYSIS BY NON LINEAR IMBEDDING

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0027

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Zucker, Steven, W.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Yale University
51 Prospect St.
New Haven, CT

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION A: Distribution approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Political science datasets contain information of interest to planners seeking to predict international relations. The goal of this projectwas to use modern data mining techniques to determine whether such data exists and, if so, to characterize it. We developed a new approach for such analysis based on geometric harmonics. At the heart of our approach is the observation that such relationships are inherently non-linear and that the data are noisy and incomplete. To demonstrate the power and usefulness of our techniques the focus was on United Nations voting data. It was shown that major historical events could be inferred from these data; that other (linear) techniques did not suffice; and that they could be extended to understanding certain aspects of international relations. We conclude that the project was successful in opening up the field of “computational international relations.”

15. SUBJECT TERMS

diffusion geometry, security, trust, political science data analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Zucker, Steven, W.
U	U	U		32	**19b. TELEPHONE NUMBER** *(Include area code)* 203 432 6434

Reset

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

DISTRIBUTION A: Distribution approved for public release.

Final Report

Steven W. Zucker
Dept. of Computer Science
Yale University
New Haven, CT 06520-8285

September 24, 2013

Abstract

Political science datasets contain information of interest to planners seeking to predict international relations. The goal of this project was to use modern data mining techniques to determine whether such data exists and, if so, to characterize it. We developed a new approach for such analysis based on geometric harmonics. At the heart of our approach is the observation that such relationships are inherently nonlinear and that the data are noisy and incomplete. To demonstrate the power and usefulness of our techniques the focus was on United Nations voting data. It was shown that major historical events could be inferred from these data; that other (linear) techniques did not suffice; and that they could be extended to understanding certain aspects of international relations. We conclude that the project was successful in opening up the field of "computational international relations."

1 Introduction

How do the religious preferences, gender, age, income and place of birth influence whether an individual will likely engage in terrorist activities? How do social and familial context affect this estimate? How does the list of countries belonging to a particular intergovernmental organization define the organization? Conversely, what does such membership imply about the country? More particularly, how might such contextual data be codified and fused into a coherent global estimate? Although the above examples are stated at different scales, they illustrate the questions facing data analysis in the social sciences. Progress on answering questions such as these from this project are reviewed below.

At a technical level, existing analysis methods, or forms of data imputation, are mainly either linear or dependent on underlyingbut unknown and perhaps unknowableprobability distributions and parameterizations. But in social situations data are rarely linearly distributed and sampling questions remain confounded. Moreover the data may be incomplete; they may be non-veridical; and they may be distorted because the subject in non-cooperative.

The approach taken in this project does not suffer from these shortcomings. It is non-linear and does not presuppose parametric forms. Instead it is based on the observation that the conceptual structure in data can often be abstracted mathematically as a low-dimensional manifold embedded in a high-dimensional space. It is based on the analysis of questionnaire data. To illustrate: each question is in effect a separate measurement; and can be considered as a separate dimension. While there may be many questions (e.g., 500 - 1000 or more) they are rarely completely independent from one another. Thus information "implicit" within the questions exists even though the subject may believe it is hidden. It is this implicit information that the subjective analysist seeks to intuit; and it is these intuitions that have driven the existing data compilations.

This approach has been applied to trade, IGO membership, conflict and voting databases. By working in collaboration with political scientists, the techniques have been refined so that it is now possible to derive embeddings representative of a number of political developments. Our experience indicates that UN voting patterns are more predictive than, e.g., IGO memberships, and that many key events, such as the development and subsequent break-up of the Soviet Union, can be readily seen. It follows, then, that there remains significant additional structure to be inferred from these databases and their integration

2 Background and Preview

Understanding the role, power, and message of InterGovernmental Organizations (IGO's) remains a key mission for political and social scientists. They provide a channel for information flow, and a source of data by which policy could be determined with which international relations (IR) could be governed. At the request of former Program Manager Dr. Lyons, this project developed specific data mining techniques to reveal the information implicit within United Nations General Assembly voting records. It was his view that understanding the structures of these networks could shed light on important hypotheses and theoretical questions in IR such as: (i) Do IGO's have any influence on armed conflicts? (ii) What do votes in the UN General Assembly reveal? (iii) Does trade between (Democratic) countries help to reduce conflicts between them? His intuition, it turns out, was largely correct.

We preview our results with Fig. 1. Although this images are small, the .pdf files can be enlarged on your screen. But more interestingly, it is helpful to view the results rotating in 3-D; we have provided a web page for these to be viewed at

`http://www.cs.yale.edu/homes/vision/zucker/embeddings.html`

The idea behind our approach, in short, is to view countries as points in a kind-of galaxy of other countries. The galaxy is arranged by a similarity measure, in this case based on UN voting patterns. Intuitively, each country is modeled by the vector is its votes in every issue. Each vote can be thought of as a kind of coordinate, with possibilities for vote 1 being YES, NO, Abstain. Vote 2 is another coordinate, perpendicular to the first, vote 3 another again perpendicular to both, and so on through the nearly 1,000 votes taken.

Of course, viewing points in 1000-dimensional spaces is impossible, nor is there that much information available. So the goal is to reduce the dimension of the space, while leaving the essential arrangement of the countries (the galaxy) effectively intact. This is done with a technique called diffusion geometry, and it is based on the idea that countries are close (in the galaxy) when they share lots of political, social, trade and economic capital. (All of these concepts are developed more fully in the body of this proposal.)

Fig. 1 shows this dimenion-reduced galaxy. This example is chosen to illustrate how our "history independent" techniques can infer major historical events just from the UN voting data, in this case France's self-isolation under de Gaulle's presidency. In 1957 France (cyan star, upper left corner) was close to the USA, UK, Belgium, Luxembourg (blue markers) in the galaxy of countries. By 1959, France, under the influend of Charles de Gaulle and his policies, began to withdraw from NATO military commands. The process was completed in 1966. Thus, when we look at the maps as time proceeds, we see France slowly move to the edge of the (blue) Western group in 1960, gradually edging further away by 1963 and planting itself in a distant position from that of the West in 1967. French foreign policy returned to the West after de Gaulle left office in 1969; notice France (cyan star, bottom left) moving back toward integration in NATO, its position in 1972-1973 got closer and closer to that of UKG (blue triangle, top left) (FRN opened up from its self-isolation, allowing UKG to join EC in 1973). Many more examples and discussion of the distance measure are contained in Sec. 6.

2.1 Overview of Final Report

An overview of the report is as follows. To start, we review diffusion geometry, a non-linear dimensionality reduction technique based on the concept of diffusion distance, which considers not only direct dyadic connections between social actors, but also all indirect paths of diffusion through intermediate neighbors. This is important in political science because influence accumulates in a manner than is not revealed by linear techniques.

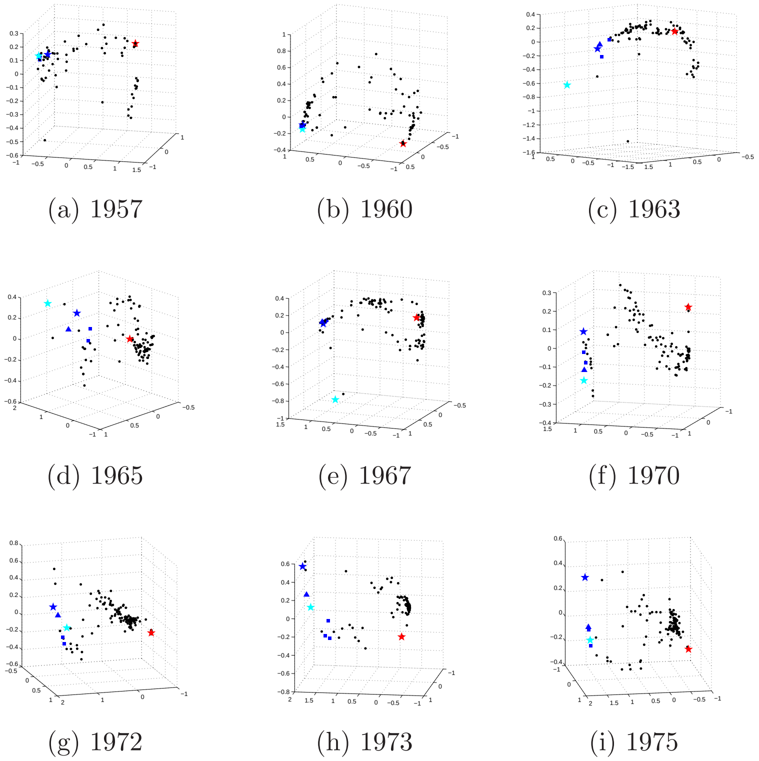
This technique has been applied to socio-political databases, such as IGO membership [13], UN voting [16]. These are described next. While these databases have received significant attention from scholars of international relations [6, 7, 10, 14] we do not believe that they have previously been analyzed by techniques such as ours. Several papers do, in effect, support our approach ([2, 8, 11, 17].

Following this, we review a hierarical clustering algorithm that was developed to identify themes running through the voting patterns. This is, in effect, the complement to the above, because it reveals structure among resolutions rather than countries. Taken together both techniques reveal how much structure is implicit within UN General Assembly voting patterns.

3 Diffusion distance

We approach the dimensionality reduction problem by means of a social network model: Consider $G(V, W)$ as a network whose vertices $i \in V$ are the countries and kernel function W_{ij}, derived from X, measures the similarity between countries i and j.

Social phenomena and trade, unlike geography, follow a different distance



(a) 1957

(b) 1960

(c) 1963

(d) 1965

(e) 1967

(f) 1970

(g) 1972

(h) 1973

(i) 1975

Figure 1: *De Gaulle's France: Diffusion maps of UN voting pattern 1957-1975. Several countries are marked for case study identification:* ★*(USA),* ▲ *(UKG),* ★ *(FRN),* ■ *(BEL, LUX, GFR),* ★ *(RUS). These maps show France started out close to the Allies in 1957. Then in 1960, France, under de Gaulle's presidency, distanced itsef from the West. The 70s saw France coming back toward the Western fold, once de Gaulle had left.*

measure. Goods and social capital *diffuse* from one place to another, perhaps through an intermediate country. Thus nearby countries matter more than distant ones. Since classical techniques preserve all pairwise Euclidean distances between the data points, we argue that not all distances should be preserved uniformly. Instead, *only short distances shoud be maintained, and even attenuated in order to preserve the local structure, while long distances should not be considered for keeping.* The argument is illustrated in Fig. 2. In political terms, we see a polarization in which two camps (B, C) closely communicate, but (A, B) barely interact with each other except through intermediary contacts located in the middle tunnel. An embedding which highlights this polarization should tighten the clusters' girth (thus *attenuating short distances*) and stretch the tunnel's length (thus *loosening long distances* and separating the two clusters from each other). Those are the characteristics of *diffusion.* While distance could have been derived from gravitational potentials in [15], to our knowledge this is the first application of diffusion distance to sociopolitical questions.

Think of a substance (e.g. money, population, or political influence) diffusing from a source point out to its neighboring points in amounts proportional to the neighbors' similarity to the source. The substance continues to diffuse to the neighbors of those neighbors, etc. Assuming a fixed amount of substance in the network, we can define $p_t(k|i)$ as the density of substance, originating from source point i, at point k at time t. Thus $p_t(k|i)$ would be high if there are many paths of length $\leq t$ connecting i to k, and low otherwise. If we take point $i = B$ on the right of Fig. 2 as the source, after t time steps, most of the substance originated from B should end up at points like $k = C$ on the right cluster, and only a small fraction ends up at points like $k = A$ on the left, because there are significantly more paths from B to C than to A. The intuitive *diffusion distance* [3] between any two points i and j is a weighted difference between the two probability density functions:

$$\begin{aligned} D_t^2(i,j) &= \|p_t(k|i) - p_t(k|j)\|_\omega^2 \\ &= \sum_k (p_t(k|i) - p_t(k|j))^2 \omega(k) \end{aligned} \tag{1}$$

where $\omega(\bullet)$ is the weight function that normalizes the distance according to the density estimate of each vertex.

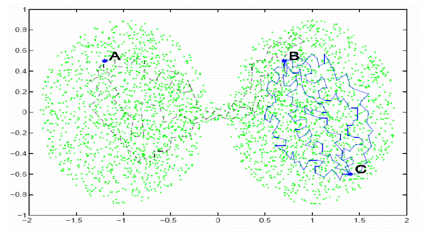


Figure 2: *Two tight clusters separated by a narrow path. It is obvious that there are many paths between any pair of nodes from the same cluster (B and C), while there are significantly fewer paths between any pair of nodes from different clusters (A and B).*

International trade can also be viewed as a diffusion process in which money diffuses from country to country. The polarization in Fig. 2 can be described in terms of trade during the Cold War. Assuming the trade pattern stays constant, the money will diffuse out to the two sources' trading partners, like 'bumps' of heat diffusing through a graph. Thus $p_t(\bullet|USA)$ will be high in the West, and low in the East, while $p_t(\bullet|USSR)$ behaves in the opposite direction. The function $p_t(\bullet|USA)$ provides a notion of "trading sphere" of the USA. Therefore, the diffusion distance between the USA and the USSR can be defined as the difference between their corresponding spheres $p_t(\bullet|USA)$ and $p_t(\bullet|USSR)$, as described by Eq. 1.

4 Random walk

In order to compute the diffusion distance $D^t(i, j)$, which takes into account *all paths* (of length t) between i and j, we begin by considering a random walk of a traveler in a network of countries $G(V, W)$. The transition probability is given by

$$M = D^{-1}W \tag{2}$$

where D is a diagonal matrix $D_{ii} = d_i = \sum_j W_{ij}$, called the degree matrix. The matrix $\widetilde{M} = D^{1/2}MD^{-1/2} = D^{-1/2}WD^{-1/2}$ is thus symmetric and has the same spectrum as M. If $p_t(i)$ denotes the probability the traveler appears in country i at time t, then

$$p_{t+1}^T = p_t^T M = p_t^T D^{-1}W \tag{3}$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of $\widetilde{M}$ and $\{\upsilon_k\}$ their corresponding orthonormal eigenvectors:

$$\widetilde{M} = \Upsilon \Lambda \Upsilon^T \tag{4}$$

where Λ is the diagonal matrix with $\{\lambda_k\}$ on its diagonal, and Υ is a matrix whose columns are the corresponding eigenvectors $\{\upsilon_k\}$.

Therefore

$$M = D^{-1/2} \widetilde{M} D^{1/2} = D^{-1/2} \Upsilon \Lambda \Upsilon^T D^{1/2} = \Psi \Lambda \Phi^T \tag{5}$$

where

$$\begin{aligned} \phi_k &= D^{1/2} \upsilon_k \\ \psi_k &= D^{-1/2} \upsilon_k \end{aligned} \tag{6}$$

which implies that $\{\phi_k\}$ and $\{\psi_k\}$ defined in Eq. 6 are the left and right eigenvectors of M corresponding to eigenvalues $\{\lambda_k\}$. Since $\{\upsilon_k\}$ are orthonormal vectors, ϕ_i and ψ_j are bi-orthonormal:

$$\phi_i^T \psi_j = \delta_{ij} \tag{7}$$

We can also verify that

$$\begin{aligned} \widetilde{M} d^{1/2} &= D^{-1/2} W D^{-1/2} d^{1/2} \\ &= D^{-1/2} W \mathbb{1} \\ &= D^{-1/2} d = d^{1/2} \end{aligned} \tag{8}$$

Therefore $d^{1/2}$ is an eigenvector of $\widetilde{M}$ with eigenvalue 1, and hence $\forall k \ |\lambda_k| \leq 1$ [12]. Thus $\lambda_1 = 1$. In fact, if G is connected (so that M represents an irreducible and aperiodic Markov chain) then $\forall k > 1 \ |\lambda_k| < 1 = \lambda_1$. We also have $\upsilon_1 = \frac{d^{1/2}}{\|d^{1/2}\|}$, which leads to $\phi_1 = \frac{d}{\|d^{1/2}\|}$ and $\psi_1 = \frac{\mathbf{1}}{\|d^{1/2}\|}$. That means ψ_1 is a constant vector, while $\phi_1(i) = \frac{d_i}{\sqrt{\sum_k d_k}}$.

Let $p_t(j|i)$ be the probability that the traveler starts walking from country i and appears in country j at time t, then it follows from Eq. 3:

$$p_t(j|i) = e_i^T M^t = e_i^T \Psi \Lambda^t \Phi^T = \sum_k \psi_k(i) \lambda_k^t \phi_k(j) \tag{9}$$

where e_i is a vector whose entry $e_i(k) = \delta_{ik}$. Therefore, if G is connected, the following limit holds, regardless of the initial starting point:

$$lim_{t\to\infty} p_t(j|i) = \psi_1(i)\phi_1(j) = \frac{d_j}{\|d^{1/2}\|^2} = \frac{d_j}{\sum_k d_k} \tag{10}$$

The first eigenvector ϕ_1 serves as the stationary distribution of the random walk M. It can also be considered a density estimate, which tells us of how frequently our walker passes by a particular country. In social network terminology, it is the centrality vector.

5 Diffusion Maps

For each country i, we can imagine the diffusion process starts with an initial distribution $p_0(j|i) = \delta_{ij}$. After t steps, this distribution diffuses out to the neighborhood of i, with the landscape described by $p_t(j|i)$. The walker is more likely to end up in states close to i than those far away. The diffusion distance $D_t^2(i,j)$ can be measured by Eq. 1, with the weight function $\omega(k) = \frac{1}{d_k}$ which normalize the distance by the centrality measure of each node. $D_t^2(i,j)$ can be seen as the weighted difference between the two distributions of concentrations after t steps of two random walks starting from nodes i and j.

We also define diffusion map Ψ_t as the mapping between the original data space onto the first κ left eigenvectors of M:

$$\Psi_t(i) = (\lambda_1^t\psi_1(i), \lambda_2^t\psi_2(i), \ldots, \lambda_\kappa^t\psi_\kappa(i)) \tag{11}$$

It is easily verifiable that the diffusion distance in Eq. 1 is equal to Euclidean distance in the diffusion map space:

$$
\begin{aligned}
D_t^2(i,j) &= \sum_l^n \Big(\sum_k^\kappa \lambda_k^t (\psi_k(i) - \psi_k(j))\phi_k(l) \Big)^2 \frac{1}{d_l} \\
&= \sum_{k_1,k_2}^\kappa \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \\
&\sum_l^n \frac{\phi_{k_1}(l)\phi_{k_2}(l)}{d_l} \\
&= \sum_{k_1,k_2}^\kappa \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \\
&\sum_l^n \frac{d_l \psi_{k_1}(l)\phi_{k_2}(l)}{d_l} \qquad (12) \\
&= \sum_{k_1,k_2}^\kappa \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \\
&\sum_l^n \psi_{k_1}(l)\phi_{k_2}(l) \\
&= \sum_{k_1,k_2}^\kappa \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \delta_{k_1 k_2} \\
&= \sum_k^\kappa \lambda_k^{2t} \Big(\psi_k(i) - \psi_k(j) \Big)^2 \\
&= \| \Psi_t(i) - \Psi_t(j) \|^2
\end{aligned}
$$

Practically, only the last $(\kappa - 1)$ coordinates are to be considered because ψ_1 is a constant vector. Additionally, since $\forall k \; |\lambda_k| <= 1$, components $\lambda_k^t \psi_k(i)$ in Eq. 11 corresponding to smaller values of λ_k vanish rapidly as t increases, achieving nonlinear dimensionality reduction.

6 Experimental Results

We present several examples of the application of our diffusion maps algorithm on geopolitical databases. Three-dimensional visualizations of the re-

sults are also made available online at http://www.cs.yale.edu/homes/vision/zucker/embeddings.ht
.

6.1 Geographical map: A physical perspective

Fig. 3 provides an experiment with geographical embedding of national capitals [5], with the kernel $W_{ij} = e^{-\frac{r_{ij}^2}{10^8}}$. The resulting embedding approximates global positions.

6.2 Intergovernmental organization (IGO) membership pattern

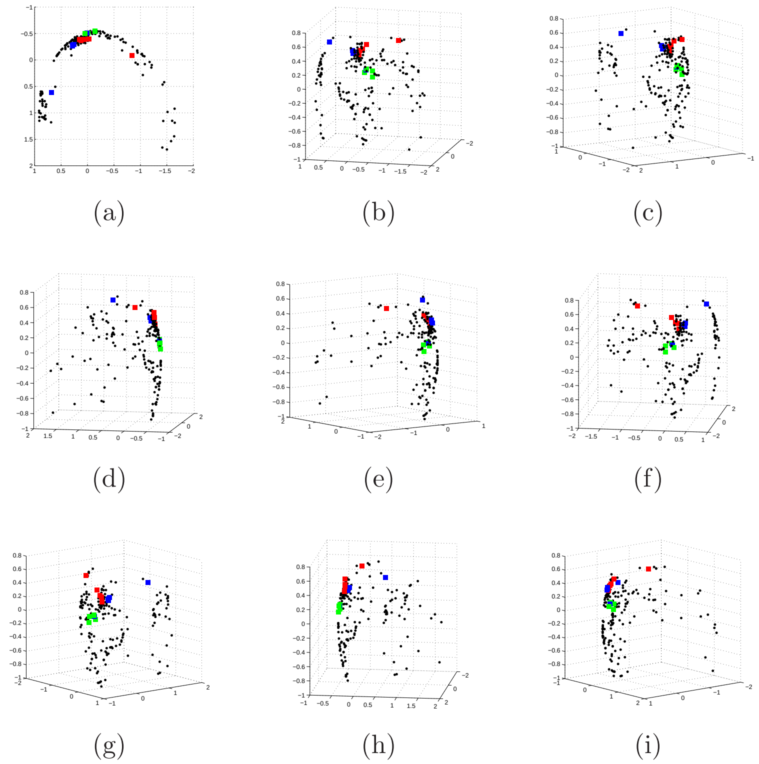
Inter-governmental organizations (IGO) play a crucial role in international relations. Fig. 4 reveals how various countries are positioned, given their IGO memberships [13] in the year 2000. The diffusion maps were derived using the correlation of joint membership as the kernel function [9]. The maps show that IGO membership pattern tends to correlate with regional geographical positions.

6.3 UN vote pattern: de Gaulle's France

Using the Pearson product correlation kernel [9], we embed the UN member nations in a three-dimensional space, according to their votes in the UN General Assembly in various years. Fig. 5 shows the embedding of the network of UN Assembly members according to their voting patterns at various time during 1957-1975. These visualizations provide us with a novel historical perspective.

Additionally, Fig. 6 plots the ratios of embedding distance in the period 1965-2000:

- $\frac{\text{dist(FRN,EU*)}}{\text{diam(EU*)}}$ as the blue line
- $\frac{\text{dist(UKG,EU*)}}{\text{diam(EU*)}}$ as the red line
- $\frac{\text{dist(FRN,UKG)}}{\text{diam(EU*)}}$ as the green line

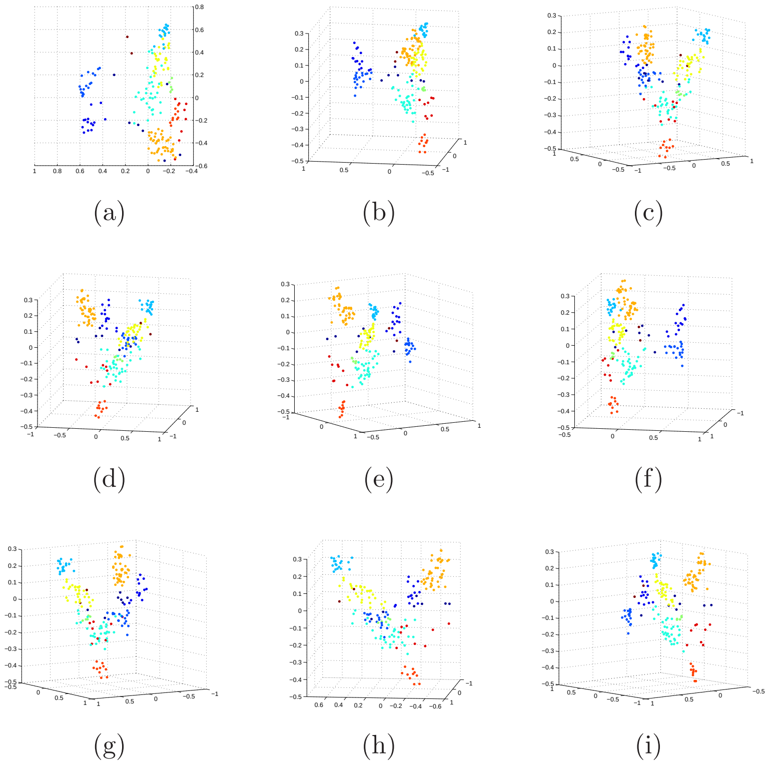


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: *Geographical embedding of national capitals in 3-dimensional space, using the* 2^{nd}*,* 3^{rd}*, and* 4^{th} *vectors of the diffusion map. The edge weight function is defined as* $W_{ij} = e^{-\frac{r_{ij}^2}{10^8}}$ *where* r_{ij} *is geographical distance between captitals of nations* i *and* j*. Figure (a) provides a top down view, while (b)-(i) show side views of the embedding from different angles, turning from west to east (counterclockwise). Several countries are marked with colored squares for easy identification:* 🟦 *(USA, UKG, FRN, BEL, ISR),* 🟥 *(RUS, CHN, POL, HUN, BLR),* 🟩 *(EGY, SYR, LEB, SAU, KUW).*



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4: *Diffusion map of countries, given their IGO membership in 2000, using the* 2^{nd}, 3^{rd}, *and* 4^{th} *vectors. (a) provides a top-down perspective, while (b)-(i) show side views from different angles, in counterclockwise rotation. The countries are manually colored according to their geographical locations, which shows again that IGO's aligning influence is mostly regional. Legend (with respect to (a)): Caribean (dark blue, upper left); Central & South American (medium blue, lower left); Western European (light blue, upper right); former Soviet states & ISR (yellow, upper right); North African (light red, middle far right); African (light orange, lower right); Middle East (dark orange, middle right); USA & CAN (dark red, middle).*

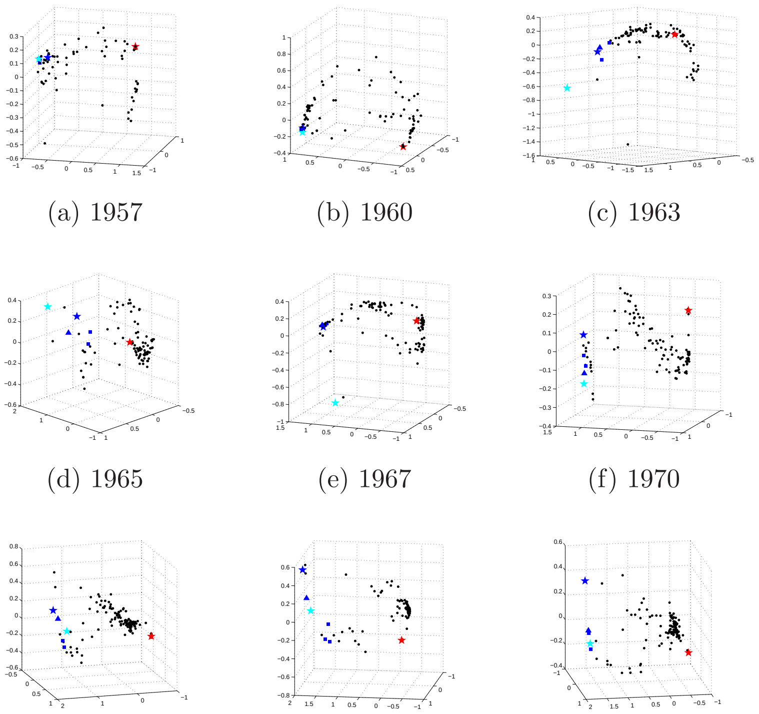
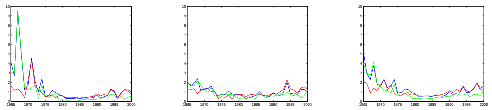


Figure 5: *De Gaulle's France: Diffusion maps of UN voting pattern 1957-1975. Several countries are marked for case study identification:* ★*(USA),* ▲ *(UKG),* ★ *(FRN),* ■ *(BEL, LUX, GFR),* ★ *(RUS). These maps show France started out close to the Allies in 1957. Then in 1960, France, under de Gaulle's presidency, distanced itsef from the West. The 70s saw France coming back toward the Western fold, once de Gaulle had left.*



(a) Diffusion distance (b) Distance of PCA embedding (c) Hamming distance

Figure 6: *Ratios of embedding distances between FRN-EU* (blue), UKG-EU* (red), FRN-UKG (green) in 1965-2000. Here EU* is defined as the states of the European Community, excluding FRN & UK. Thes plots show how relations between France, UK and the rest of the Western European states changed over time, with France standing far apart during the 60s, and coming back to the fold afterward.*

where EU* is defined as the states of the European Community, excluding FRN & UK. The distances and diameters are calculated from diffusion distance, distance of PCA embedding, and Hamming distance of the VOTE matrix. The plots of different distance measures show us how diffusion method amplifies the connections between highly connected actors, and also enhances separation between distant parties.

France's self-isolation under de Gaulle's presidency is apparent from the diffusion maps. In 1957 (Fig. 5a), France (cyan star, upper left corner) was close to the USA, UK, Belgium, Luxembourg (blue markers). By 1959, France under Charles de Gaulle began to withdraw from NATO military commands and completed that process in 1966. Thus, when we look at the maps as time proceeds, we see France slowly move to the edge of the (blue) Western group in 1960 (Fig. 5b), gradually edging further away by 1963 (Fig. 5c), planting itself in a distant position from that of the West in 1967 (Fig. 5e). The distance ratio plot in Fig. 6a shows us the blue line (FRN-EU) started at around 0.8, the green line (FRN-UKG) reaching its peak at 9 in 1967-1968, while the red line (UKG-EU) lying low initially, indicating France's isolated position from that of the Western countries (and UKG) at the time. After de Gaulle left office in 1969, we see the blue line begin to decline steeply, moving in tandem with the red line, implying a reverse course in France' foreign policy, gradually edging closer to that of the rest of West. Indeed, Fig. 5f shows France (cyan star, bottom left) moving back toward integration in NATO, its position in 1972-1973 (Fig. 5g-5h) got closer and closer to that of UKG (blue triangle, top left) (FRN opened up from its self-isolation, allowing UKG to join EC in 1973). By 1975 (Fig. 5i) France again stood close to the Western bloc. In the 80s until the end of the Cold War, the distance ratios FRN-EU and UKG-EU (blue & red lines, Fig. 6a) ascended slightly, due to the absorption of new members into the EU. The green line (FRNK-UKG), however, remains low throughout the 80s, showing how close FRN and UKG's policies were to each other during that period.

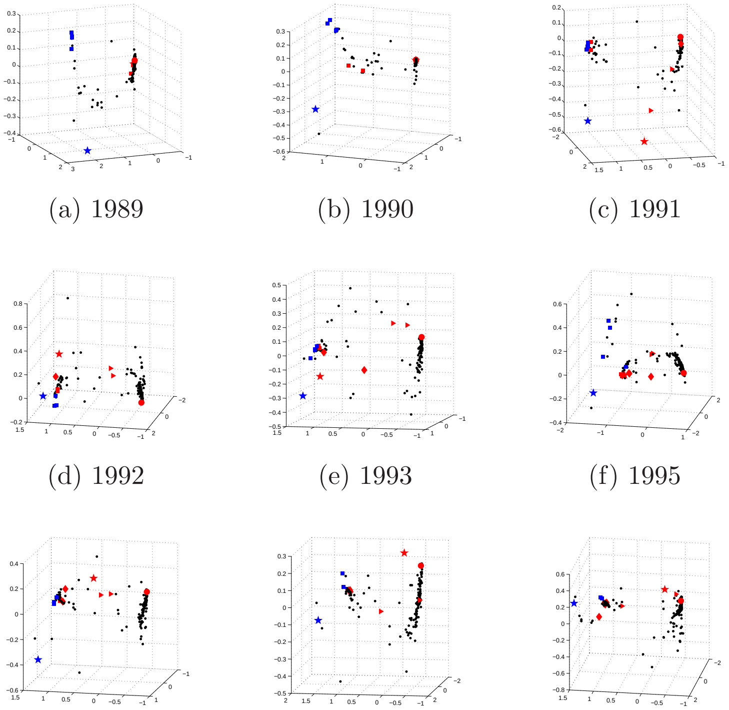
The diffusion maps reveal the inherently low dimensional structure among countries, in agreement with prior analysis [1, 7]. It is also apparent from Fig. 6 that PCA fails to discover a pattern in the movements of countries in the network, while diffusion distance uncovers the same pattern as the simple Hamming distance. The spectrum given by PCA decays very slowly: it requires 20-30 dimensions to describe all variances in the voting data. Diffusion method, on the other hand, requires only 5-7 dimensions to describe the voting patterns [7]. The diffusion method performs better in amplifying significant events in its distance plot (e.g. the period from 1957-1967 in which France isolated itself). However, the diffusion distance in Fig. 6 is computed from only 5 dimensions, whereas the Hamming distance is the aggregated result of votes on all UN resolutions in a particular year.

6.4 UN vote pattern: The collapse of the Soviet Union

Fig. 7 shows the maps of nations according to their UN voting patterns at various time during 1989-2005. The embedded positions are computed by our diffusion method such that countries are placed closer to each other if they voted similarly, and far apart if they did not. Fig. 8 compares 3 distance metrics: (a) diffusion distance by our method (which shall be defined in more details later in this article), (b) PCA embedded distance (Euclidean distance between data points embedded by a Principal Component Analysis projection), and (c) Hamming distance (normalized number of resolutions that countries voted different from each other.) Each subfigure plots the ratios of embedding distances in the period 1965-2000:

- $\frac{\text{dist(USA,EU)}}{\text{diam(EU)}}$ as the blue line
- $\frac{\text{dist(RUS,EU)}}{\text{diam(EU)}}$ as the red line
- $\frac{\text{dist(POL,EU)}}{\text{diam(EU)}}$ as the green line

where EU is defined as the states of the European Community.



(a) 1989

(b) 1990

(c) 1991

(d) 1992

(e) 1993

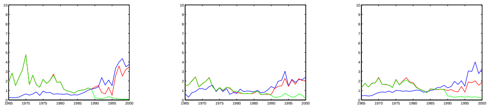
(f) 1995

(g) 1997

(h) 2000

(i) 2005

Figure 7: *The collapse of the Soviet Union: Diffusion maps of UN voting pattern 1989-2005. Several countries are marked for case study identification:* ★ *(USA),* ■ *(UKG, FRN, BEL, LUX),* ★ *(RUS),* ♦ *(YUG),* ▶ *(UKR, BLR),* ■ *(POL, HUN),* ● *(CHN).*



(a) Diffusion distance

(b) Distance of PCA embedding

(c) Hamming distance

Figure 8: *Ratios of embedding distances between USA-EU (blue), RUS-EU (red), POL-EU (green) in 1965-2000. Here EU is defined as the states of the European Community. Thes plots show how relations between USA, USSR, Poland and the Western European states changed over time, with Poland tailing the USSR until 1989, after which it was completely aligned with the West.*

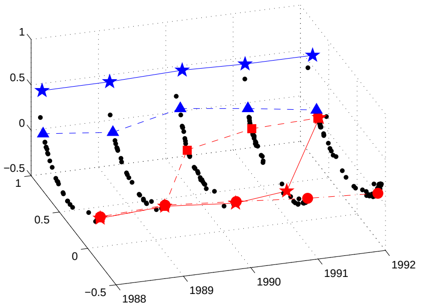


Figure 9: *The disintegration of the Soviet Union (1988-1992): The evolution of 2-dimensional diffusion maps of nations according to their voting patterns in the UN Assembly. Each dot denotes the global position of a country in a particular year. Special markers are drawn to denote:* ★ *(USA),* ▲ *(UKG),* ★ *(RUS),* ■ *(POL),* ● *(CHN). Several lines are also plotted connecting the "paths" of these countries over time. Note how USA and UKG stayed relatively steady at their positions, while the paths of Communist states started to diverge since 1989. POL was the first to move out of the camp in 1990, followed by RUS, whereas CHN remained in their original position throughout the whole period.*

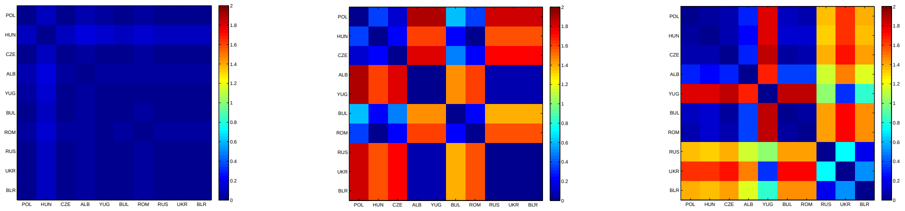
The 1989 diffusion map is polarized with the Western bloc (blue) on the left and the Eastern bloc (red) on the right of Fig. 7a. The distance ratio plots in Fig. 8a clearly shows the green line (POL-EU) trailing the red line (RUS-EU) prior to 1989, indicating Poland's policy completely dominated by that of the Soviet Union. However, in 1990 (Fig. 7a), Poland and Hungary (red squares) switched to the left, followed quickly by Czecholovakia, Bulgaria, and then the three newly independent Baltic republics. Fig. 8a clearly reveals a break between the green line and the red line from 1989, showing different trends in Poland and Russia's policies from then on. By 1991 (Fig. 7b), Russia (red star), Belarus, and Ukraine (2 red triangles) followed suit, as they moved toward the center. In 1992, after the Soviet bloc fully disintegrated (Fig. 7d), its members had all migrated to the left, with Ukraine and Belarus hanging in the middle, leaving China (red circle) on the right, close to the Arabs and the third world. Figs. 7d- 7f depicts Russia's effort to get close to the West, as Yeltsin vied for Western support for admission to NATO or the EU. The downward trend of the red line during 1992-1995 in Fig. 8a indicates Russia's aborted attempt to get close to the EU. After Yeltsin's second election in 1996 and his failure to court the West (Fig. 7g), Russia moved to the right of the map. Fig. 8a records a sharp ascent of the red line after 1996, implying Russia's abandonment of its westward movement. Further shift eastward occurred after Putin replaced Yeltsin in 2000 (Fig. 7h), as Russia switched to the right, getting close to China again.

The collapse is even more evident in Fig. 9, which provides a time-evolution of the event by stringing the 2-dimensional structures of the alignments in Fig. 7 along the time dimension. It is apparent from the figure:

- USA and UKG stood close to each other in the 2-dimensional alignment, and their distance remain relatively stable throughout the 5-year period.

- The break-up of the Soviet Union is shown in the diverging lines of RUS, POL and CHN. The Union stayed intact until 1990, when POL moved away, toward the other side of the map. In 1991, RUS inched apart from CHN and the third-world countries, and then moved completely out by 1992.

For further analysis, we consider the group of Communist countries in the years 1989-1991. Fig. 10 shows the diffusion distances among these countries



(a) 1989

(b) 1990

(c) 1991

Figure 10: *Diffusion distances among the countries in the Communist Bloc (POL, HUN, CZE, ALB, YUG, BUL, ROM, RUS, UKR, BLR) in 1989-1991. The colors denote distance value from low (cool, blue color) to high (hot, red color).*

in 1989-1991. The group was tight in 1989 and quickly disintegrated in 1990 and 1991, as the diffusion distances suddenly spiked up in these two years.

7 Themes across Resolutions

We now switch emphasis to inferring implicit structure among resolutions. Since voting patterns are responsible for the global embedding, further insight can be obtained by looking at those resolutions that have the highest variance among clusters of countries. In essence we are asking: among nearby countries, which topics are most controversial; i.e., on which neighbors vote differently. We focus, in particular, on the Soviet bloc of Eastern European countries.

Numerical values are assigned to votes:

$$\begin{array}{ccc} \text{against} & \rightarrow & \text{-1} \\ \text{abstain} & \rightarrow & 0 \\ \text{for} & \rightarrow & +1. \end{array}$$

so we can compute the variances of the votes of the Eastern Bloc for every UN resolution in the three years around the breakup of the Bloc.

Table 1 shows a topical breakdown of the 20 highest-variance resolutions among these countries votes in 1989-1991. During the first year most of the attention remained focused on old Cold War issues and matters of development, anti-colonialism, and human rights in the global south on which the Soviet bloc had commonly sided with less developed countries against the

	1989	1990	1991
Middle East	2	7	6
Weapon Nonproliferation	2	6	5
Anti-Apartheid & Human Rights	6	2	2
Territory & Sovereignty	5	5	6
Others	5	0	1

Table 1: *Topical breakdown of the 20 highest-variance resolutions according to the votes of Eastern Bloc members (POL, HUN, CZE, ALB, YUG, BUL, ROM, RUS, UKR, BLR) during 1989-1991.*

developed north and west. But even as soon as 1990 and then 1991 those divisive issues had faded, and in their place Middle Eastern issues, especially focusing on Israel and the Palestinians became dominant. On those issues the US and Israel were in a minority even among other western states. Consequently they became, and have remained, apart from the Assembly majority as they had never been before.

It is clear from this example that there are currents in the resolutions. Our next goal is to discover them automatically. In order to not have preconceived notions, we adapt a hierarchical clustering algorithm and an eigenfunction summary method next.

8 Building hierarchical clustering trees

We now seek to organize the resolutions according to how countries voted on them, with the goal of uncovering themes that summarize them. Given the lack of a prior on themes among resolutions – how many there are or, even, whether any exist – we adapt a hierarical clustering algorithm.

For each cluster in the hierarchy, we seek a set of "summary questions" that best approximate large groups of questions underlying the embeddings. This has two advantages: (i) it reduces the dimension of the data set; and (ii) if the summary questions are combinations of small numbers of questions, they are simple to interpret. This latter point is an advantage to the political scientists. We stress that our approach is in contrast to factor analysis, which leads to factors that are linear combinations of all questions.

Any pair of resolutions are related if they are either highly correlated or highly anticorrelated. For example, during the Cold War period, a UN resolution condemning Israel in Middle East issues will most likely be rejected by the West and supported by the Arabs; however, another UN resolution in support of Israel would lead to the exact opposite voting pattern. Therefore, we study the absolute value of data correlation as a topical similarity function.

More formally, this leads to a relatively standard objective function that only depends on dot products. It can be modified using the kernel trick to incorporate non-linearities, in particular those that arise with our diffusion kernel.

We treat each resolution as a vector of responses $\boldsymbol{q}_i$ normalized so

$$\sum_j \boldsymbol{q}_i(j) = 0$$

$$\|\boldsymbol{q}_i\| = 1$$

We denote $Q = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$, the set of votes to all resolutions.

On the way to designing an objective function, we first seek to find a set of "summary questions" $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s_k}\}$ and a clustering $C = \{c_1, \ldots, c_k\}$ of questions with summary questions with the following properties:

$$\bigcup_{i=1}^{k} c_i = Q \tag{13}$$

$$c_i \cap c_j = \varnothing, i \neq j \tag{14}$$

$$\|\boldsymbol{s}_i\| = 1 \tag{15}$$

Equations (13) and (14) make sure that each question is assigned to a single cluster. We now want to maximize the similarity between each question and the summary question it is assigned to. The objective function we seek to maximize is definied as:

$$\phi(C, S) = \sum_{i=1}^{k} \sum_{\boldsymbol{q_j} \in c_i} |\langle \boldsymbol{q}_j | \boldsymbol{s}_i \rangle|^2$$

In the bioinformatics community this objective is called the diametric clustering objective function [4]. This has an equivalent metric clustering minimization problem. Using the fact that $|\langle \boldsymbol{q}_j | \boldsymbol{s}_i \rangle|^2 \leq 1$

$$
\begin{aligned}
\arg\max_{C,S} \phi(C,S) &= \arg\min_{C,S} \{n - \phi(C,S)\} \\
&= \arg\min_{C,S} \sum_{i=1}^{k} \sum_{\boldsymbol{q_j} \in c_i} d(q_j, s_i)^2
\end{aligned}
$$

where $d(\boldsymbol{v},\boldsymbol{w}) = \sqrt{1 - |\langle \boldsymbol{v}|\boldsymbol{w}\rangle|^2}$
$d(\cdot,\cdot)$ is a pseudometric, which is to say

1. $d(\boldsymbol{v},\boldsymbol{v}) = 0$

2. $d(\boldsymbol{v},\boldsymbol{w}) = d(\boldsymbol{w},\boldsymbol{v})$

3. $d(\boldsymbol{u},\boldsymbol{v}) + d(\boldsymbol{v},\boldsymbol{w}) \geq d(\boldsymbol{u},\boldsymbol{w})$

1 and 2 are trivial. Proof of 3 is technical, and is omitted for space reasons.

The maximization version of this problem suggests one simple heuristic, while the minimization problem suggests another. The first is a modification of Lloyd's algorithm.

```
procedure ModifiedLloyd($\{q_1, \ldots, q_n\}$)
    cluster = initialclustering()
    while $\phi_{old} \neq \phi_{new}$ do
        $\phi_{old} = \phi_{new}$
        for $i = 1$ to $k$ do
            $V = concat(q \in c_i)$                ▷ $V = [q_{c1}|\cdots|q_{cm}]$
            $v_i = SVD(V)$                         ▷ $v_i$ is largest left sing. vect.
        end for
    end while
    for $j = 1$ to $n$ do
        put $q_j$ in the cluster that maximizes $|\langle v_i|q_j\rangle|$
    end for
    recompute $\phi_{new}$
end procedure
```

This algorithm increases the objective function ϕ at each stage. In fact, each for loop increases ϕ.

It is instructive to consider how this might be proved. The second for loop is straightforward, as each question is assigned to the cluster that maximizes the objective. Therefore if any questions change cluster, the objective function will increase.

Let V be defined as above. Then $V = U * D * W^T$ where U and V are unitary and D is a diagonal matrix of singular vectors. Then

$$\begin{aligned}\sum_{\boldsymbol{q_j} \in c_i} |\langle \boldsymbol{q}_j | \boldsymbol{s} \rangle|^2 &= \|\boldsymbol{s}^T V\|^2 \\ &= \sum_i D_{ii}^2 \langle \boldsymbol{u_i} | \boldsymbol{s} \rangle\end{aligned}$$

where $\boldsymbol{u_i}$ are the columns of U. This is maximized by setting $\boldsymbol{s}$ to be equal to the largest singular vector $\boldsymbol{u}_1$

Therefore each stage of the algorithm increases ϕ. Since there are a finite number of clusterings, and hence values for ϕ and each stage of the algorithm increases ϕ, it converges, though possibly not to the global optimum.

8.1 Toward Thematic Hierarical Clustering

Although Lloyd's algorithm guarantees a local maximum in the objective function, for our application we seek a related – but in a local sense, slightly different – condition: we guarantee that the absolute correlation distance cannot exceed a threshold. Guaranteeing this condition was deemed a necessity by the political scientists, and leads to a variation on the above algorithm.

We start with n individiual singleton clusters of entities E and a data matrix D of m countries (rows) and n resolutions (columns), such one shown in Table 2. We also have a correlation threshold $\theta \in (0, 1)$ and a cooldown ratio $\alpha \in (0, 1)$. We repeatedly iterate through the following steps, merging clusters until only one remains:

```
procedure GreedyCluster(D, $\theta$)
    unallocated = D
    for c in unallocated do
        remove c from unallocated
        for q in unallocated do
            if $abs(corr(c, q) < \theta)$ then
                remove q from unallocated
```

	#3508	#3510	#3515	#3538	#3570
USA	-1	-1	-1	-1	-1
UKG	-1	-1	-1	0	0
RUS	1	1	1	1	1
POL	0	0	0	0	0
CHN	1	1	0	1	1

Table 2: *An excerpt from the UN voting data [16] of 5 countries (USA, UKG, RUS, POL, CHN) in 1990 on 5 issues, denoted by their roll call id's (RCID): #3508 (Dissemination of information on decolonization) #3510 (Observer status of national liberation movements recognized by the OAU and/or by the League of Arab States) #3515 (Cessation of all nuclear test explosions) #3538 (Calls upon Israel to become party to the Treaty on the Non-Proliferation of Nuclear Weapons) #3570 (Status of the International Convention on the Suppression and Punishment of the crime of Apartheid). The votes are represented by numbers: 1 (Yes), 0 (Abstain), -1 (No).*

```
                assign q to cluster c
            end if
        end for
    end for
    reassign questions to most correlated cluster center
    return clusters
end procedure
procedure GreedyTree(D, $\theta$, $\alpha$)
    while numclusters $> 1$ do
        clusters = GreedyCluster(D, $\theta$)
        set D to largest singular vector of each cluster
        $\theta = \theta\alpha$
    end while
end procedure
```

Performance is very similiar to the Lloyd algorithm, which could in effect be inserted into the first procedure.

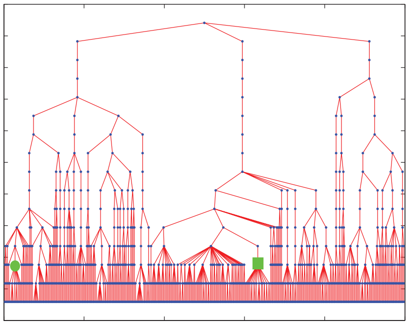


Figure 11: Clustering result of UN resolutions during the period 1998-2002. Two individual clusters are marked with ● and 🟩 symbols for demonstration.

8.2 Results on UN Resolutions

We applied the clustering algrithm on the set of UN resolutions during the period 1998-2002 [16], with $\theta = 0.95$ and $\alpha = 0.8$. Fig. 11 shows the clustering hierarchy with two clusters ● and 🟩. The resolutions in cluster ● pertain only to Middle East-related resolutions, while cluster 🟩 comprises resolutions from two topics (Human Rights and Nuclear Disarmaments).

We take a more detailed look at the resolutions during the breakup of the Soviet bloc of countries in Figs.12 - 14.

9 Summary

In this project we developed a diffusion-based approach to embedding high-dimensional UN voting data and showed how to cluster the resolutions "driving" these embeddings. Organization among countries revealed political relationship, and cluster analysis revealed thematic threads running across time. In effect we showed that much of the historical record can be "read out" from UN voting patterns.

10 Publications Arising from this Project

- Liberty, E., and Zucker, S.W., The Mailman algorithm for matrix vector multiplication, *Information Processing Letters*, 2009, **109**(3), 179 -

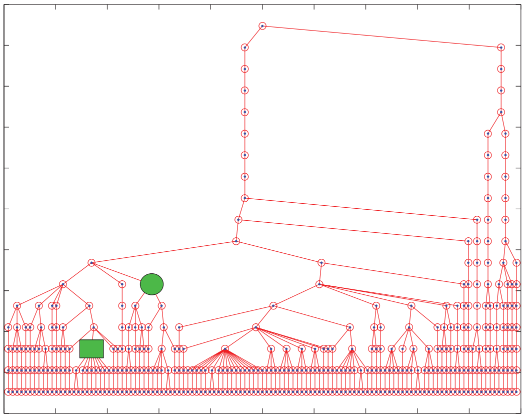


Figure 12: *Thematic clustering of UN Resolutions 1989. The* 🟩 *cluster is about Middle East issues, while the* ● *is about disarmament and nuclear weapons. The variance in voting patters across Eastern Bloc countries on these issues is virtually 0.*

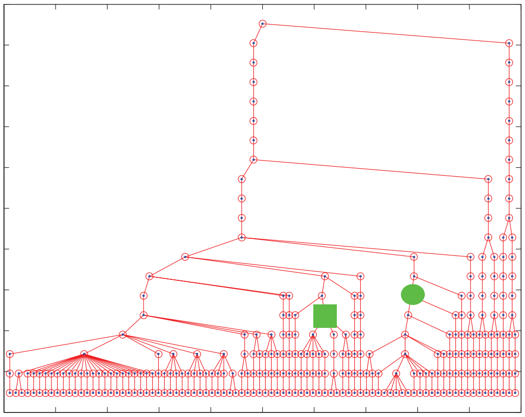


Figure 13: *Thematic clustering of UN Resolutions 1990. The variance across clusters starts to increase, indicating political change. The cluster* ● *on Middle East issues is growing larger, while others (e.g.* 🟩 *remain fixed on nuclear weapons issues.*

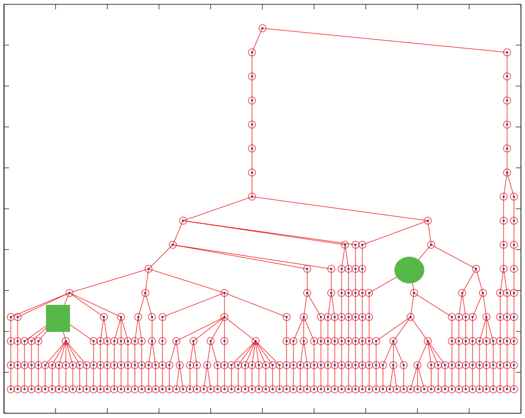


Figure 14: *Thematic clustering of UN Resolutions 1991. Again the Middle East* ● *cluster remains while the nuclear weapons cluster* 🟩 *enlarges to include economic and other testing issues.*

182.

- Keller, Y., Lafon, S., Coifman, R., and Zucker, S.W. Audio-Visual Group Recognition by diffusion maps, *IEEE Transactions On Signal Processing*, 2010, **58**(1), 403 - 413.

- Le, Minh Tam, Sweeney, J., Liberty, E. and Zucker, S.W., Similarity Kernels via Bi-Clustering for Conventional Intergovernmental Organizations, Current Issues in Predictive Approaches to Intelligence and Security Analytics (PAISA-10), *IEEE Intelligence and Security Informatics Conference*, Vancouver, 26 May, 2010.

- Zucker, S. W., Geoemtric Harmonics Reveal Information Structure, *Networks and Networking in the Humanities*, National Endowment for the Humanities and Insitute for Pure and Applied Mathematics, UCLA, August 12, 2010.

- Le, Minh-Tam, Russett, B., Sweeney, J., and Zucker, S.W. A Kernel Analysis of Kant's Conjecture, *4th Annual Political Networks Conference*, University of Michigan, Gerald R. Ford School of Public Policy, June 14-18, 2011.

- Le, Minh-Tam, Sweeney, J., Russett, B., and Zucker, S.W., Structural Inference in Political Science Datasets. *Proc. IEEE ISI Intelligence* *and Security Informatics*, Washington, DC, 11 - 14 June, 2012.

- Minh-Tam Le, John Sweeney, Matthew Lawlor and Steven W. Zucker, Discovering Thematic Structure in Political Datasets, *Proc. IEEE ISI Intelligence and Security Informatics*, June 4-7, 2013, Seattle, WA.

- A web site was developed to introduce and illustrate these concepts; see: `http://www.cs.yale.edu/homes/vision/zucker/embeddings.html`.

References

[1] H. R. Alker. Dimensions of conflict in the general assembly. *The American Political Science Review*, 58(3):642–657, 1964.

[2] H.R. Alker and B. Russett. Discovering voting groups in the general assembly. *American Political Science Review*, 60(1):327–339, 1996.

[3] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[4] Inderjit S Dhillon, Edward M Marcotte, and Usman Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.

[5] K. S. Gleditsch and M. D. Ward. Measuring space: A minimum-distance database and applications to international studies. *J. of Peace Research*, 38(6):739–758, 2001.

[6] H. Hegre, J. R. Oneal, and B. Russett. Trade does promote peace: New simultaneous estimates of the reciprocal effects of trade and conflict. *J. of Peace Research*, 47(6):763–774, 2010.

[7] S. Y. Kim and B. M. Russett. Voting alignments in the general assembly. In B. M. Russett, editor, *The Once and Future Security Council*, pages 29–57. New York: St. Martin's Press, 1997.

[8] B. Kinne. Multilateral trade and militarized conflict: Centrality, opennness, and asymmetry in the global trade network. *Journal of Politics*, 74:308–322, 2012.

[9] M.-T. Le, J. Sweeney, B.M. Russett, and S.W. Zucker. Structural inference in political science datasets. In *Proc. of IEEE ISI*, pages 138–140, 2012.

[10] A. G. Long. Bilateral trade in the shadow of armed conflict. *ISQ*, 52:81–101, 2008.

[11] Z. Maoz. *Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816-2001.* New York: Cambridge University Press, 2011.

[12] C. D. Mayer. *Matrix Analysis and Applied Linear Algebra*, chapter 8, pages 661–704. SIAM, 2000.

[13] J. C. Pevehouse, T. Nordstrom, and K. Warnke. The COW-2 international organizations dataset version 2.0. *Conflict Management and Peace Science*, 21(2):101–119, 2004.

[14] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand. A network analysis of committees in the u.s. house of representatives. *PNAS*, 102(20):7057–7062, 2005.

[15] J. Q. Stewart. Demographic gravitation: Evidence and applications. *Sociometry*, 11(1/2):31–58, 1948.

[16] A. Strezhnev and E. Voeten. United nations general assembly voting data, 2012.

[17] E. Voeten. Clashes in the assembly. *International Organization*, 54(2):185–215, 2000.

AFOSR Deliverables Submission Survey

Response ID:2870 Data

1.

AFOSR

TODAY'S BREAKTHROUGH SCIENCE FOR TOMORROW'S AIR FORCE

Report Submission Form

surveygizmo

If you have any questions, please contact your Program Manager or Assistant Program Manager.

Air Force Office of Science and Research
875 Randolph Street
Suite 325 Room 3112
Arlington, VA 22203

1. Report Type

Final Report

Primary Contact E-mail
Contact email if there is a problem with the report.

steven.zucker@yale.edu

Primary Contact Phone Number
Contact phone number if there is a problem with the report

203-432-6434

Organization / Institution name

Yale University

Award Information

Grant/Contract Title
The full title of the funded effort.

Social Data Analysis by Non Linear Embedding

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-09-1-0027

Principal Investigator Name
The full name of the principal investigator on the grant or contract.

Zucker, Steven, W.

Program Manager
The AFOSR Program Manager currently assigned to the award

Benjamin Knott

Report Information - Annual Report

Report Information - Final Report

Report Information - Conference/Workshop Report

Report Information - Equipment Report

Report Information - Patent/Invention Report, DD882

Report Information - Financial Report, SF425

Report Information - STTR Status Report

Report Information - STTR Annual Progress Report

For an annual report, the reporting period start date is either start date of the grant, if this is the first report, or 1 day after the due date of the previous report. The end date is due date of this report.

The reporting period start and end dates are the start and end dates of the award.

Reporting Period Start Date

12/15/2008

Reporting Period End Date

09/24/2013

Report Abstract:

In the Abstract section, please list any accomplishments that have been made since the last report submission (or since the beginning of the award if this is the first report).
Please do not type "see report" here, include at least an abstract, 250 words or more, of the accomplishments mentioned in your report.

Report Abstract:

In the Abstract section, enter the Final Conference Report. This is a summary of all scientific papers presented and a list of all attendees.

Report Abstract:

In the Abstract section, enter the Final Performance Report.

The Final Performance Report will identify the acquired equipment (although it may vary from that described in your proposal) by name and associated costs. The Final Performance Report shall summarize the research or educational project for which the equipment will be used.

The patent and inventions coverage contained in Article 36, Intangible Property, of the Research Terms and Conditions does not apply to this award.

Article 15, Intangible Property, in the AFOSR Agency Specific Requirements

does not apply to this award.

Abstract

Political science datasets contain information of interest to planners seeking to predict international relations. The goal of this projectwas to use modern data mining techniques to determine whether such data exists and, if so, to characterize it. We developed a new approach for such analysis based on geometric harmonics. At the heart of our approach is the observation that such relationships are inherently nonlinear and that the data are noisy and incomplete. To demonstrate the power and usefulness of our techniques the focus was on United Nations voting data. It was shown that major historical events could be inferred from these data; that other (linear) techniques did not suffice; and that they could be extended to understanding certain aspects of international relations. We conclude that the project was successful in opening up the field of "computational international relations."

Distribution Statement
This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement
If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

NOTE: Extra documentation is NOT required for this report. If you would like to send additional documentation, send it directly to your Program Manager or Assistant Program Manager.

SF298 Form
Please attach your SF298 form. A blank SF298 can be found here. Please do not spend extra effort to password protect or secure the PDF, we want to read your SF298. The maximum file size for SF298's is 50MB.

af-cover.pdf

Upload the Report Document. The maximum file size for the Report Document is 50MB.

final-report.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Additional Information

Archival Publications (published) during reporting period:

Liberty, E., and Zucker, S.W., The Mailman algorithm for matrix vector multiplication, Information Processing Letters, 2009, 109(3), 179 - 182

Keller, Y., Lafon, S., Coifman, R., and Zucker, S.W. Audio-Visual Group Recognition by diffusion maps, IEEE Transactions On Signal Processing, 2010, 58(1), 403 - 413.

• Le, Minh Tam, Sweeney, J., Liberty, E. and Zucker, S.W., Similarity Kernels via Bi-Clustering for Conventional Intergovernmental Organizations, Current Issues in Predictive Approaches to Intelligence and Security Analytics (PAISA-10), IEEE Intelligence and Security Informatics Conference, Vancouver, 26 May, 2010.

• Zucker, S. W., Geoemtric Harmonics Reveal Information Structure, Networks and Networking in the Humanities, National Endowment for the Humanities and Insitute for Pure and Applied Mathematics, UCLA, August 12, 2010.

• Le, Minh-Tam, Russett, B., Sweeney, J., and Zucker, S.W. A Kernel Analysis of Kant's Conjecture, 4th Annual Political Networks Conference, University of Michigan, Gerald R. Ford School of Public Policy, June 14-18, 2011.

• Le, Minh-Tam, Sweeney, J., Russett, B., and Zucker, S.W., Structural Inference in Political Science Datasets. Proc. IEEE ISI Intelligence and Security Informatics, Washington, DC, 11 - 14 June, 2012.
• Minh-Tam Le, John Sweeney, Matthew Lawlor and Steven W. Zucker, Discovering Thematic Structure in Political Datasets, Proc. IEEE ISI

Intelligence and Security Informatics, June 4-7, 2013, Seattle, WA.

- A web site was developed to introduce and illustrate these concepts; see: http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

Changes in research objectives (if any):

none

Change in AFOSR Program Manager, if any:

Original Program Manager: Terrence Lyons

Second Program Manager: Joseph B. Lyons

Current Program Manager: Benjamin Knott

Extensions granted or milestones slipped, if any:

P00005

For an STTR Status or STTR Annual Progress Report, please e-mail your program manager directly.

2. Thank You

E-mail user

Sep 24, 2013 11:33:08 Success: Email Sent to: steven.zucker@yale.edu

Your report has been submitted. You should receive an email confirmation soon that it is being processed by AFOSR. Please print this page as proof of submission. Thank you.

Principal Investigator Name:	Zucker, Steven, W.
Primary Contact E-mail:	steven.zucker@yale.edu
Primary Contat Phone Number:	203-432-6434
Grant/Contract Title:	Social Data Analysis by Non Linear Embedding
Grant/Contract Number:	FA9550-09-1-0027
Program Manager:	Benjamin Knott
Report Type:	Final Technical
Reporting Period Start Date:	12/15/2008
Reporting Period End Date:	09/24/2013
Abstract:	Political science datasets contain information of interest to planners seeking to predict international relations. The goal of this projectwas to use modern data mining techniques to determine whether such data exists and, if so, to characterize it. We developed a new approach for such analysis based on geometric harmonics. At the heart of our approach is the observation that such relationships are inherently non- linear and that the data are noisy and incomplete. To demonstrate the power and usefulness of our techniques the focus was on United Nations voting data. It was shown that major historical events could be inferred from these data; that other (linear) techniques did not suffice; and that they could be extended to understanding certain aspects of international relations. We conclude that the project was successful in opening up the field of “computational international relations.”
Distribution Statement:	Distribution A - Approved for Public Release
SF298 Form:	55-c3b619c750c74f728661bddafd359a83_af-cover.pdf
Report Document	13-fd71672bd6ab3c31406faf3a2972c159_final-report.pdf
Archival Publications:	Liberty, E., and Zucker, S.W., The Mailman algorithm for matrix vector multiplication, Information Processing Letters, 2009, 109(3), 179 - 182 Keller, Y., Lafon, S., Coifman, R., and Zucker, S.W. Audio-Visual Group Recognition by diffusion maps, IEEE Transactions On Signal Processing, 2010, 58(1), 403 - 413.

- Le, Minh Tam, Sweeney, J., Liberty, E. and Zucker, S.W., Similarity Kernels via Bi-Clustering for Conventional Intergovernmental Organizations, Current Issues in Predictive Approaches to Intelligence and Security Analytics (PAISA-10), IEEE Intelligence and Security Informatics Conference, Vancouver, 26 May, 2010.

- Zucker, S. W., Geoemtric Harmonics Reveal Information Structure, Networks and Networking in the Humanities, National Endowment for the Humanities and Insitute for Pure and Applied Mathematics, UCLA, August 12, 2010.

- Le, Minh-Tam, Russett, B., Sweeney, J., and Zucker, S.W. A Kernel Analysis of Kant's Conjecture, 4th Annual Political Networks Conference, University of Michigan, Gerald R. Ford School of Public Policy, June 14-18, 2011.

- Le, Minh-Tam, Sweeney, J., Russett, B., and Zucker, S.W., Structural Inference in Political Science Datasets. Proc. IEEE ISI Intelligence and Security Informatics, Washington, DC, 11 - 14 June, 2012.
- Minh-Tam Le, John Sweeney, Matthew Lawlor and Steven W. Zucker, Discovering Thematic Structure in Political Datasets, Proc. IEEE ISI Intelligence and Security Informatics, June 4-7, 2013, Seattle, WA.

- A web site was developed to introduce and illustrate these concepts; see: http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

Changes in Research objectives: none

Change in AFOSR Program Manager, if any: Original Program Manager: Terrence Lyons

Second Program Manager: Joseph B. Lyons

Current Program Manager: Benjamin Knott

Extensions granted or milestones slipped, if any: P00005

Response ID: 2870

Survey Submitted:	Sep 24, 2013 (11:33 AM)
IP Address:	130.132.173.250
Language:	English (en-US,en;q=0.5)
User Agent:	Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:23.0) Gecko/20100101 Firefox/23.0
Http Referrer:	
Page Path:	1 : (SKU: 1) 2 : Thank You (SKU: 2)
SessionID:	1380035592_5241ac089efcb2.09096491

Response Location

Country:	United States
Region:	CT
City:	New Haven
Postal Code:	06520
Long & Lat:	Lat: 41.308102, Long:-72.9282